1 A few remarks on cross-validation for Lasso or Ridge

Let us consider the Lasso problem:

$$\underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \mathcal{L}_{lasso}(\theta, \alpha) := \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \alpha ||\theta||_1.$$

For a given α, the standard procedure would be to learn θ on a train sample $(y_i, x_i)_{i=1}^n$ and then to assess the predictive quality of the model learnt on a test sample $(\widetilde{y}_i, \widetilde{x}_i)_{i=1}^m$. We denote by $\widehat{\theta}(\alpha)$ the estimated parameter.

$\widehat{\theta}(\alpha)$ may vary a lot for different values of α and thus the predictive quality of the model may also change a lot for different αs. How can we choose this hyperparameter α? We can use what is called the K-fold cross validation procedure. It takes the following form

1. split randomly the train set in K folds of approximately the same size. These folds form a partition $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ of the training set.

2. choose a set of possible values of α (on a grid) that we want to try. We denote this set $\mathcal{A}$. In practice, $\mathcal{A}$ has to be a finite set.

3. for every $\alpha \in \mathcal{A}$, we operate the following loop

   (a) for every $k \in \{1, \dots, K\}$, we compute $\widehat{\theta}_k(\alpha)$ using all the training set but $\mathcal{I}_k$.

   (b) we use $\mathcal{I}_k$ as a test set on which we compute the MSE associated with $\widehat{\theta}_k(\alpha)$, that we call $\text{MSE}\big(\widehat{\theta}_k(\alpha)\big)$.

   $\Longrightarrow$ at the end of the loop, we take $\text{Score}(\alpha) := \frac{1}{K} \sum_{k=1}^{K} \text{MSE}\big(\widehat{\theta}_k(\alpha)\big)$ as a measure of the performance of α.

4. we select $\alpha_* \in \operatorname{argmin}_{\alpha \in \mathcal{A}} \text{Score}(\alpha)$.

5. we use this α_* as our final choice. We learn $\widehat{\theta}(\alpha_*)$ on the whole training set and evaluate the performance on the test set.

Note that we could use other criteria than MSE to choose α_* (the Mean Absolute Error for instance). To further reduce the risk of overfitting, we could split the data in three subsamples: one to choose α_* (on which we do steps 1 to 4 of the cross-validation algorithm mentioned above), one to learn $\widehat{\theta}(\alpha_*)$ and one to evaluate the performance of the algorithm.

The cross-validation approach is exactly the same for the Ridge estimator.